Biochem. J. (1991) 274, 481–490 (Printed in Great Britain)

Characterization and organization of the genes encoding the A-, B- and C-chains of human complement subcomponent C1q

The complete derived amino acid sequence of human C1q

Grant C. SELLAR,*‡ Derek J. BLAKE†§ and Kenneth B. M. REID*‖

*M.R.C. Immunochemistry Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, and † Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, U.K.

A partial cDNA clone for the A-chain of human complement subcomponent C1q was isolated from a monocyte library. Use of the A-chain cDNA clone, and a previously characterized B-chain cDNA clone [Reid (1985) Biochem. J. **231**, 729–735] allowed the isolation of overlapping cosmid clones that were shown to contain the genes encoding the A-, B- and C-chains of human C1q. The three genes were found to be aligned, 5′→3′, in the same orientation, in the order A–C–B on a 24 kb stretch of DNA on chromosome 1p. The A-, B- and C-chain genes are approx. 2.5, 2.6 and 3.2 kb long respectively, and each contains one intron, located within a codon for a glycine residue found half-way along the collagen-like region present in each chain. These glycine residues are located just before the point where the triple-helical portions of the C1q molecule appear to bend when viewed in the electron microscope. Southern-blot analyses indicated that there is only one gene per chain, and preliminary examination of genomic DNA from several C1q-deficient patients showed no evidence for major deletions or insertions within the A-, B- or C-chain genes. The DNA sequence of the coding region of the C-chain gene allows the completion of the entire derived amino acid sequence for the human C1q molecule. The globular, *C*-terminal, regions of the chains of C1q show a strong similarity in amino acid sequence to the non-collagen-like, *C*-terminal, regions of the type VIII and type X collagens, indicating structural and evolutionary relationships between these three molecules.

INTRODUCTION

Complement subcomponent C1q (460 kDa) associates with the proenzymes C1r and C1s (found as a Ca^{2+}-dependent 360 kDa $C1r_2C1s_2$ complex) to yield C1, the first component of the serum complement system. Electron-microscopy studies have shown C1q to be composed of six globular heads linked via six collagen-like stalks to a fibril-like central region (Svehag *et al.*, 1972; Shelton *et al.*, 1972; Knobel *et al.*, 1975; Brodsky-Doyle *et al.*, 1976). The C1q molecule contains 18 polypeptide chains (six A-, six B- and six C-chains), each of which is approx. 225 amino acid residues long and contains a collagen-like region of approx. 81 residues, located near the *N*-terminus, and a *C*-terminal globular region of approx. 136 residues (Reid, 1983). The hexameric structure, seen in the electron microscope, is due to the formation of six triple-helical regions, between the collagen-like sequences present in the six A-, six B- and six C-chains, which are viewed as the central region and stalks, and each of the globular heads is formed by association of the *C*-terminal region of an A-, a B- and a C-chain (Reid, 1976; Reid & Porter, 1976). The collagen-like regions of C1q interact with the Ca^{2+}-dependent $C1r_2C1s_2$ proenzyme complex (Reid *et al.*, 1977; Siegel & Schumaker, 1983), and efficient activation of C1 takes place on interaction of the globular heads of C1q with the Fc regions of IgG or IgM antibody present in immune complexes. The entire amino acid sequences of the A- and B-chains and the collagen-like region of the C-chain have been established by previous protein studies (Reid & Thompson, 1978; Reid, 1979; Reid *et al.*, 1982) and cDNA studies (Reid, 1985). In the present paper we describe the characterization and sequencing of genomic clones encoding the A- and C-chain genes, which provide an ordering of the A-, B- and C-chain gene cluster, on chromosome 1p, and derivation of the complete amino acid sequence of the human C1q molecule. Determination of the entire C1q sequence allows comparison of all three chains with the chains of the type VIII (Yamaguchi *et al.*, 1989) and type X (Ninomiya *et al.*, 1986) collagens. Sequence similarity between the *C*-terminal globular regions of C1q and the *C*-terminal regions of these collagens indicates an evolutionary relationship between these three molecules. The chains of several carbohydrate-binding proteins, such as mannan-binding protein (Drickamer *et al.*, 1986; Taylor *et al.*, 1989; Lu *et al.*, 1990), lung surfactant protein SP-A (White *et al.*, 1985; Voss *et al.*, 1988) and conglutinin (Davis & Lachman, 1984; Strang *et al.*, 1986), also contain a mixture of collagen-like and globular sequences (which are clearly different from those found in C1q). These lectins show a strong structural similarity to C1q when viewed in the electron microscope and one of them, mannan-binding protein, has been shown to mimic C1q functional activity by its ability to activate the $C1r_2C1s_2$ proenzyme complex after interaction with suitable carbohydrate ligands (Lu *et al.*, 1990).

MATERIALS AND METHODS

Screening of cDNA and genomic libraries

A human monocyte λgt 10 cDNA library was purchased from Clontech (Cambridge BioScience, Cambridge, U.K.). The library

‡ Present address: The Children's Hospital, Division of Immunology, Boston, MA 02115, U.S.A.
§ Present address: Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, U.K.
‖ To whom correspondence and requests for reprints should be addressed.
The nucleotide sequence data reported will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases.

had an average insert size of 0.8 kb and a complexity of 1.4×10^5 clones before amplification. Approx. 2.5×10^5 plaques were screened with a 49-base-long unique oligonucleotide probe corresponding to amino acid residues 109–124 of the A-chain of C1q (Reid *et al.*, 1982) and constructed by using the codon bias shown by the B-chain cDNA data (Reid, 1985). Rescreening of positive clones was carried out with a 53-base-long unique oligonucleotide probe and a mixture of 64 17-base-long oligonucleotide probes corresponding to residues 186–203 and residues 188–193 of the A-chain respectively. The oligonucleotide probes were synthesized on an Applied Biosystems (Warrington, Cheshire, U.K.) DNA Synthesiser, according to the manufacturer's instructions, deprotected by incubation at 55 °C for 14 h in conc. NH_3 solution (2.0 ml), then purified by ethanol precipitation and electrophoresis on a polyacrylamide gel. The oligonucleotide probes were radiolabelled and used for hybridization as described by Hanahan & Meselson (1983).

Approx. 5×10^5 plaques of a λ '4X' human genomic library were screened with a 0.6 kb A-chain cDNA probe isolated from the monocyte cDNA library. The DNA used to construct the λ '4X' genomic library was prepared from the cell line GM1416B (48; XXXX) derived from a patient tetraploid with respect to the X-chromosome and diploid with respect to autosomes. The DNA was partially digested with the restriction enzyme *Mbo*I, size-selected to give an average insert size of 15–23 kb, and cloned into the λ vector EMBL3.

Approx. 2×10^5 cosmid clones from a human '4X' cosmid library were screened with the 0.6 kb A-chain cDNA probe and a 0.9 kb B-chain cDNA probe (Reid, 1985). The DNA used to prepare the cosmid library was from the same '4X' cell line used in the construction of the λ '4X' EMBL3 library. For the cosmid library, the DNA was partially digested with *Mbo*I, and fragments, of 35–45 kb, were size-fractionated by sucrose-density-gradient centrifugation and ligated into the cosmid vector pDVcosA2 (Knott *et al.*, 1988). The cDNA and genomic libraries were screened by using standard procedures (Maniatis *et al.*, 1982; Frischauf *et al.*, 1983; Glover, 1985). Nucleotide sequencing was performed by both the chemical degradation method (Maxam & Gilbert, 1980) and the dideoxy chain-termination method (Sanger *et al.*, 1977). Sequenase (U.S. Biochemical Corp., Cleveland, OH, U.S.A.) was used instead of DNA polymerase 1 Klenow fragment for the dideoxy chain-termination method and the primers used in the reactions were either the M13 Universal Primer or a unique oligonucleotide directed against a region of sequenced DNA. Fragments for sequencing were subcloned into M13 mp8 or the *Bam*HI and/or *Eco*RI sites of the cloning vector pBS (KS+).

Southern blotting and Northern blotting

Southern blotting of cloned and genomic DNA was carried out as described by Southern (1979) and Wahl *et al.* (1979). Isolation of total RNA was carried out as described by Chirgwin *et al.* (1979) and Northern blotting was carried out as described by Maniatis *et al.* (1982).

Samples of DNA from families with one, or more, members showing a complete deficiency of C1q function

Family B. Genomic DNA samples from four members of Family B in which one child (female P) was C1q-deficient were provided by Dr. M. J. Walport (Hammersmith Hospital, London, U.K.). Samples of DNA were obtained from the mother (N), father (C), the patient (P) and a sister (J) but not from the third child, a brother (J). The patient's serum showed no C1q by either functional or antigenic tests and the patient exhibited systemic lupus-erythrematosus-like symptoms (from 21 years) and suffered from dementia (from 25 years), fungal skin infections and retinitis. The patient died at 28 years of age from uncontrolled bronchopneumonia.

Family EM. Samples of genomic DNA from three members (the mother, F, the father, A, and a C1q-deficient child, R, who showed lupus-like disease) of Family EM were provided by Professor G. Hauptmann (Clinique Dermatologigue et Institut d'Immunologie, Strasbourg, France). Of the six siblings in the family four are homozygous for the defect (brother R and sister G, who show a lupus-like disease, brother H, who has cutaneous lesions, and brother M, who was healthy at the time of the study), one was heterozygous for the defect and healthy (sister Z) and one was homozygous normal (sister A). The defect was characterized by the presence of an abnormal dysfunctional C1q molecule that was unable to interact with the Fc region of aggregated immunoglobulin or with the $C1r_2C1s_2$ complex (Chapuis *et al.*, 1982; Grosshans *et al.*, 1987). Low concentrations of the abnormal C1q were detected in the sera of both parents, who were healthy.

Family F. Samples of genomic DNA from five members of Family F were provided by Dr. J. B. Ziegler (The Prince of Wales Children's Hospital, Rundwick, N.S.W., Australia). Two of the three siblings (brother S and sister J) showed functional C1q deficiency, and the patient, S, had cerebral lupus but no renal disease. The third child, brother D, showed a normal C1q function and concentration.

RESULTS

Characterization of an A-chain cDNA clone

Thirty-nine duplicating positive clones were detected on screening the human monocyte λgt 10 cDNA library with the 49-base-long oligonucleotide probe. Only one of these clones rescreened positively with the 53-base-long and the mixture of 64 17-base-long oligonucleotide probes. This clone was found, following *Eco*RI digestion of the purified λ DNA, to contain an insert of 516 bp, and sequencing studies showed that it coded for the amino acid residues 111–223 of the A-chain followed (Fig. 1) by a termination codon plus 174 bp of 3′ non-coding sequence (including the putative polyadenylation sequence AATAAA, which is underlined in Fig. 1). Extensive screening of the monocyte λgt 10 cDNA library was carried out with the 516 bp A-chain cDNA probe, as well as a variety of oligonucleotide probes, but no further cDNA clones for the A-chain were detected. Two other cDNA libraries (a 3-day cultured human monocyte λgt 10 cDNA library and a human spleen λgt 10 cDNA library) were also screened in order to try and isolate a full-length A-chain cDNA clone. However, no positive clones were obtained from either of these libraries. Attempts to isolate cDNA clones for the C-chain of human C1q from these three libraries and a variety of other cDNA libraries, including the liver cDNA library from which the B-chain cDNA clone was isolated (Reid, 1985), were also unsuccessful. It was therefore decided to complete the cloning of all three chains at the genomic level.

Characterization of the A-chain gene

One positive clone (λ4XC1qA) was obtained on screening 5×10^5 clones of the λ '4X' genomic library with the 516 bp A-chain cDNA probe. The insert was estimated to be 18 kb long. An 8 kb *Eco*RI fragment, derived from the insert, which was considered would contain the complete A-chain gene, was subcloned into the *Eco*RI site of the cloning vector pBS (KS+) and designated 'pC1qA8.0E'. The nucleotide sequence obtained for the A-chain gene, from this subclone, is shown in Fig. 1, and

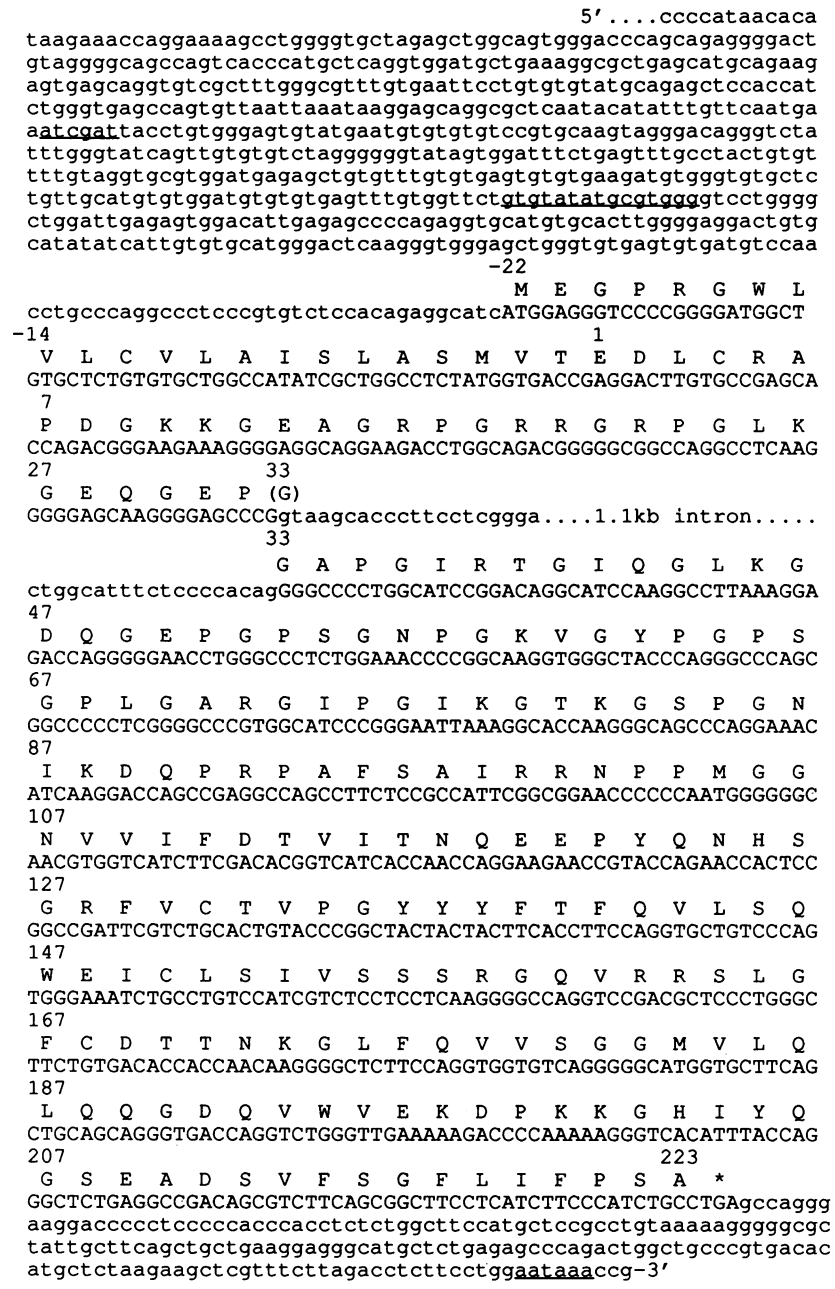

Fig. 1. Nucleotide sequence coding for the A-chain of human C1q

The derived protein sequence is shown above the nucleotide sequence and is predicted to include a 22-residue-long chain leader peptide. The 5′ limit of exon 1 of the A-chain gene was estimated to be in the region of the underlined *Cla*I site (atcgat), although a putative TATA box 180 bp downstream is also underlined. The termination codon TGA is indicated by * and the polyadenylation signal aataaa is underlined. The nucleotide sequence obtained for the 516 bp A-chain cDNA clone, corresponding to residues A111 to A223 and all the 3′ non-coding sequence, was found to be identical with that determined for the genomic clone. At positions A75, A81, A150 and A156, the derived sequence from the DNA studies indicates proline, lysine, cysteine and serine respectively rather than lysine, proline, asparagine and tryptophan as deduced from the protein sequencing (Reid, 1979; Reid *et al.*, 1982).

it indicates that the A-chain has a leader sequence of 22 amino acid residues and that this sequence plus residues 1–32 and the first base of the codon for Gly-33 are present in the first exon of the A-chain gene, which is followed by a 1.1 kb intron and a second exon of approx. 1 kb. Although the 5′ limit of exon 1 has not yet been formally established, the presence of 735 bp nucleotides of coding sequence, 180 bp of 3′ untranslated sequence and an estimated average 200-base-long poly(A) tail, in conjunction with an estimated mRNA size of 1.4 kb (see below), suggests that the transcription site is approx. 300 bp 5′ to the nucleotides coding for Met-22 (Fig. 1). Thus, with an intron of 1.1 kb, the entire A chain gene is approx. 2.5 kb long.

Characterization of cosmid clones containing the A-, B- and C-chain genes

A single cosmid clone 'cosBQ1' containing the entire B-chain gene had been characterized previously (Reid, 1985). This cosmid clone did not contain A-chain or C-chain coding sequence as

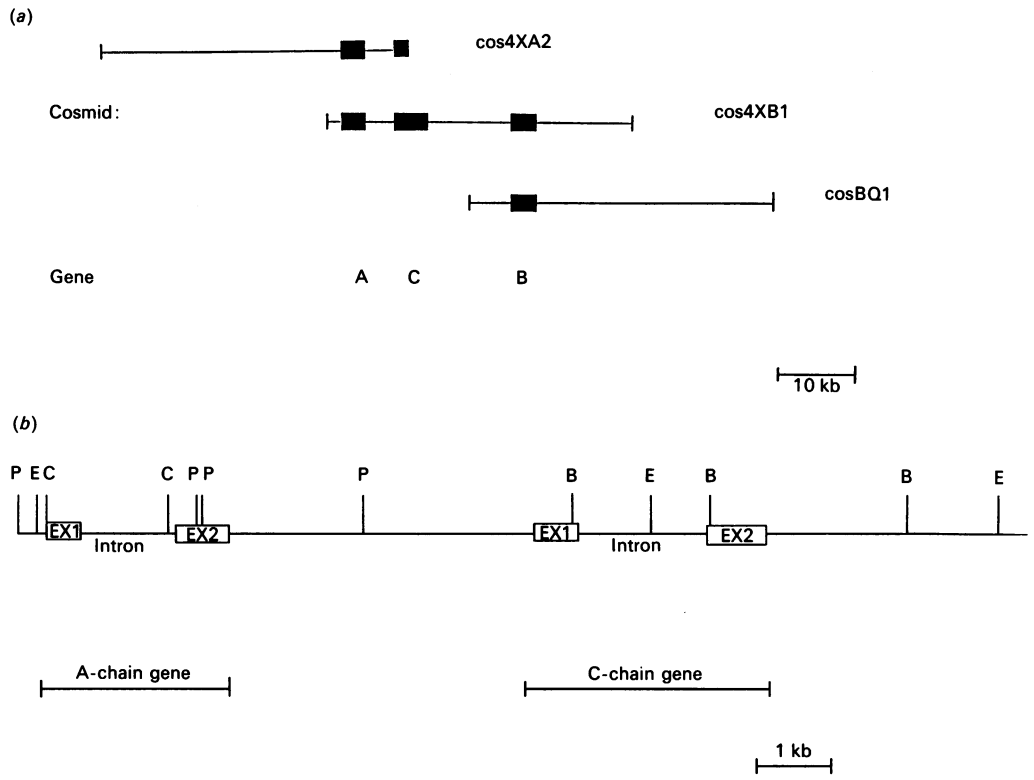

Fig. 2. (*a*) Alignment of the cosmid clones containing the A-, B- and C-chain genes for human C1q and (*b*) a limited restriction map of the C1q A- and C-chain genes

Key to restriction enzymes: B, *Bam*HI; C, *Cla*I; E, *Eco*RI; P, *Pvu*II. A restriction map of the region of the B-chain gene is given in Reid (1985).

judged by using a variety of oligonucleotide probes (constructed according to the available known amino acid sequence of the A-chain and partial sequence of the C-chain) and the A-chain 516 bp cDNA probe. Two positive clones (cos4XA2 and cos4XB1) were obtained on screening the '4X'-cosmid library with the 516 bp A-chain cDNA probe, and one of these (cos4XB1) was also positive with the 0.9 kb B-chain cDNA probe. By use of the A-chain and B-chain cDNA probes and restriction-enzyme mapping the three cosmid clones were found to overlap as shown in Fig. 2, thus encompassing approx. 80 kb of genomic DNA. The A-chain and B-chain cDNA probes were found to hybridize to restriction fragments in a manner that was consistent with the presence of the A-chain gene on both cos4XA2 and cos4XB1 and of the B-chain gene on both cos4XB1 and cosBQ1 (Fig. 2). Identification of the location of the C-chain gene was obtained when two C-chain 5′ oligonucleotide probes (based on the known *N*-terminal amino acid sequence of the C-chain) were found to hybridize to a common set of restriction fragments derived from both the cos4XA2 and cos4XB1 cosmid clones. Also, use of a 39-base-long oligonucleotide probe, based on the C-chain 'globular' sequence, indicated that the coding region for the globular portion of the C-chain was on cos4XB1 and not on cos4XA2 or cosBQ1. It was therefore concluded that clone cos4XB1 contained the genes encoding the A-, B- and C-chains of C1q (Fig. 2), and this was confirmed by sequencing the coding regions of the C-chain gene as shown in Fig. 3. The DNA sequence predicts that the C-chain has a leader sequence of 28 amino acid residues and that this sequence plus residues 1–32 and the first base of the codon for Gly-33 are present in the first exon of the C-chain gene, which is followed by a 1.7 kb intron and a second exon of approx. 900 bp. Although the 5′ limit of exon 1 has not yet been formally established, the presence of 735 nucleotide residues of coding sequence, 300 bp of 3′ untranslated sequence and an estimated 200-base-long poly(A) tail in conjunction with an estimated mRNA size of 1.5 kb (see below) indicates that the transcription site is approx. 300 bp 5′ to the nucleotide residues coding for Met-28 (Fig. 3). The entire C-chain gene is therefore estimated to be approx. 3.2 kb long. The gene cluster encoding the A-, B- and C-chains of human C1q is therefore located on a stretch of 24 kb of DNA on the short arm of chromosome 1p (in view of the A- and B-chain genes being assigned to chromosome 1p; Sellar *et al.*, 1987), with the three genes aligned head-to-tail, 5′→3′, in the order A–C–B.

Northern-blot analysis

The 516 bp A-chain cDNA probe, 0.9 kb B-chain cDNA probe and a 1.6 kb C-chain genomic probe were used, consecutively, to probe a Northern blot of polyadenylated mRNA isolated from human monocytes cultured for 7 days in 10% (v/v) autologous serum. This showed that the sizes of the A-, B- and C-chains of the mRNAs were 1.4, 1.4 and 1.5 kb respectively. Polyadenylated mRNA isolated from human non-adherent lymphocytes that had been cultured for 7 days in 10% (v/v) autologous serum gave no detectable signal on a Northern blot with the cDNA and genomic probes for the chains of human C1q.

Southern-blot analysis of DNA from C1q-deficient patients

The A- and B-chain cDNA probes and C-chain genomic probe were used in Southern-blot analysis of genomic DNA from three families in which one or more of the family members showed complete functional C1q deficiency. The probes used in the analysis were: the 516 bp A-chain cDNA (from exon 2); the 0.9 kb B-chain cDNA (covering exons 1 and 2); a 1.6 kb C-chain

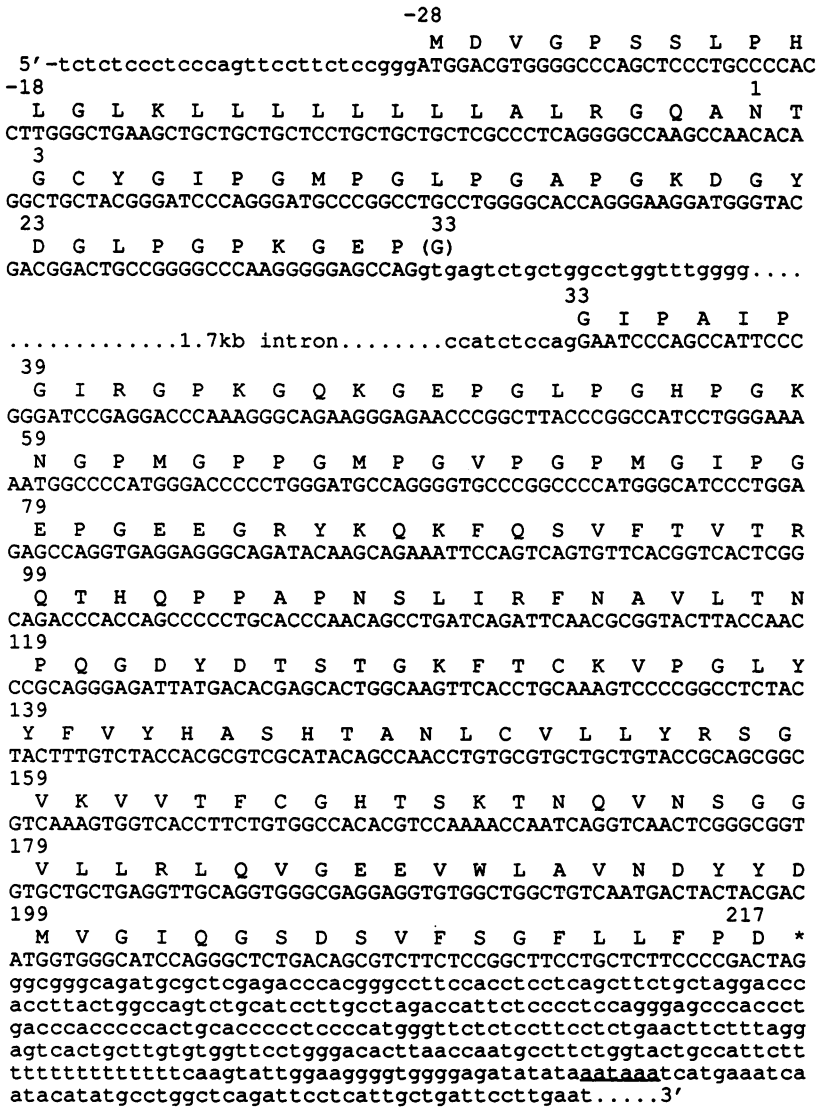

Fig. 3. Nucleotide sequence of the C1q C-chain gene

The nucleotide sequence of the human C1q C-chain gene is shown with the derived protein sequence given above. The 5′ limit of exon 1 was not determined and 28-residue-long leader peptide is predicted. The termination codon TAG is indicated by * and the putative polyadenylation signal aataaa is underlined. The derived amino acid sequence for the human C-chain shows differences on comparison with the published protein data (Reid, 1979), where at positions C29, C38, C44, C56, C59 and C62 the residues lysine, proline, lysine, proline, asparagine and methionine were predicted rather than the hydroxyproline, hydroxylysine, hydroxyproline, hydroxylysine, aspartic acid and asparagine residues identified by sequencing. It is not yet known if any of these differences represent polymorphic sites.

genomic fragment (covering exon 2). Four restriction enzymes, *Eco*RI, *Taq*I, *Bgl*II and *Pvu*II, were used in the study. In all cases no differences were seen between the Southern-blotting band patterns obtained for normal individuals and for known C1q-deficient patients and their immediate families.

DISCUSSION

The restriction mapping and sequencing studies have established that the genes encoding the chains of human C1q are arranged in a 5′→3′ orientation, in the order A–C–B, with the A-chain and C-chain genes being separated by approx. 4 kb and the C-chain and B-chain genes being separated by approx. 11 kb (Fig. 2). The cluster of C1q-chain genes has been mapped to chromosome 1p34.1–1p36.3 (Sellar & Reid, 1989), but the direction of transcription and order of the genes with respect to the centromere is as yet unknown. Two other clusters of genes for proteins associated with the complement system are known to lie on chromosome 1: (i) the genes of the regulators of complement activation (RCA) locus at 1q32, which includes the α- and β-chains of C4-binding protein, complement receptor 1, complement receptor 2, decay acceleration factor, membrane cofactor protein and Factor H (Rey-Campos *et al.*, 1988; Pardo-Manuel *et al.*, 1990); (ii) the α- and β-chains of the terminal attack component C8, within 2.5 kb of each other, in the 1p region (Kaufman *et al.*, 1989). However, the RCA locus is clearly quite distant from the genes encoding the chains of C1q, and further mapping studies are required to determine whether the C8 α-chain and β-chain genes are relatively close to the C1q genes, although lack of similarity in amino acid sequence between the chains of C8 and the chains of C1q indicates the proteins are not related.

Use of the A-, B- and C-chain probes on Northern blots of mRNA from cultured monocytes showed clearly that macro-

(a)

```
                    -28             -20                 -10               -1
Human A:                        M E G P R G W L V L C V L A I S L A S M V T
Human B:              M M M K I P W G S I P V L M L L L L L G L I D I S Q A
Human C:            M D V G P S S L P H L G L K L L L L L L L L A L R G Q A
Mouse B:                  M K T Q W G E V W T H L L L L L L G F L H V S W A
Conserved residues                                L     L
```

(b)

```
                    1                 10                  20
Human A:            E D L C R A P D G K K G E A G R P G R R
Human B:            Q L S C T G P P A I P G I P G I P G T P
Human C:            N T G C Y G I P G M P G L P G A P G K D
Mouse B:            Q S S C T G P P G I P G I P G V P G V P
Conserved residues        C         G     G         P G

                    21                30                  40
Human A:            G R P G L K G E Q - - - G E P G A P G I
Human B:            G P D G Q P G T P G I K G E K G L P G L
Human C:            G Y D G L P G P K - - - G E P G I P A I
Mouse B:            G S D G Q P G T P G I K G E K G L P G L
Conserved residues  G     G     G           G E   G   P

                    41                50                59
Human A:            R T G I Q G L K G D Q G E P G P S G N P
Human B:            A - G D H G E F G E K G D P G I P G N P
Human C:            P - G I R G P K G Q K G E P G L P G H P
Mouse B:            A - G D L G E F G E K G D P G I P G T P
Conserved residues      G     G     G     G   P G     G   P

                    60                70                  79
Human A:            G K V G Y P G P S G P L G A R G I P G I
Human B:            G K V G P K G P M G P K G G P G A P G A
Human C:            G K N G P M G P P G M P G V P G P M G I
Mouse B:            G K V G P K G P V G P K G T P G P S G P
Conserved residues  G K   G     G P   G     G     G     G

                    80                89
Human A:            K G T K G S P G N I
Human B:            P G P K G E S G D Y
Human C:            P G E P G E E G R Y
Mouse B:            R G P K G D S G D Y
Conserved residues    G     G     G
```

(c)

```
                    90                  100                 110
Human A:            K D Q P R P A F S A I R R - N P P M G G N V V I - F
Human B:            K A T Q K I A F S A T R T I N V P L R R D Q T I R F
Human C:            K Q K F Q S V F T V T R Q T H Q P P A P N S L I R F
Mouse B:            R A T Q K V A F S A L R T I N S P L R P N Q V I R F
Conserved residues                F       R                       I   F

                            120                 130                 140
Human A:            D T V I T N Q E E P Y Q N H S G R F V C T V P G Y Y
Human B:            D H V I T N M N N N Y E P R S G K F T C K V P G L Y
Human C:            N A V L T N P Q G D Y D T S T G K F T C K V P G L Y
Mouse B:            E K V I T N A N E N Y E P R N G K F T C K V P G L Y
Conserved residues      V   T N         Y         G   F   C   V P G   Y

                                    150                     160
Human A:            Y F T F Q V L S Q - W E I C L S I V S S R G Q V R
Human B:            Y F T Y H A S S R G - N L C V N L M R - G R E R A Q
Human C:            Y F V Y H A - S H T A N L C V L L Y R S G V - - - -
Mouse B:            Y F T Y H A S S R G - N L C V N L V R G R D R D S M
Conserved residues  Y F           S         L C

                              170                 180
Human A:            R - S L G F C D T T N K G L F Q V V S G G M V L Q -
Human B:            - K V V T F C D Y A Y N - T F Q V T T G G M V L K -
Human C:            - K V V T F C G H T S K T N - Q V N S G G - V L L R
Mouse B:            Q K V V T F C D Y A Q N T - F Q V T T G G V V L K -
Conserved residues            F C                 Q V     G G   V L
```

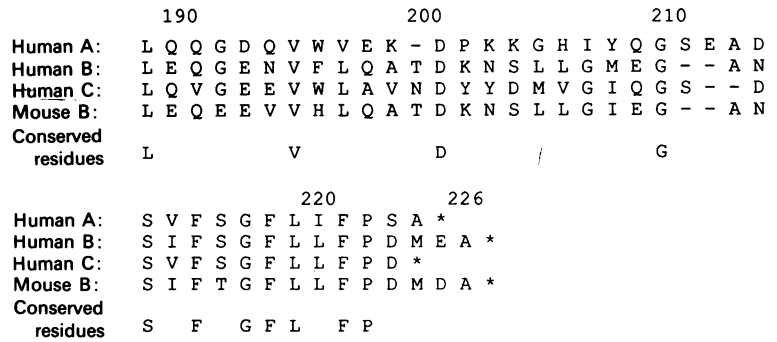

Fig. 4. Alignment of the derived amino acid sequences of the A-, B- and C-chains of human C1q and the B-chain of mouse C1q (Petry *et al.*, 1989; Wood *et al.*, 1988)

(*a*) Alignment of the leader peptides. (*b*) Alignment of the *N*-terminal collagen-like regions. – denotes that a gap has been inserted. The human B-chain numbering is used. The lysine and proline residues in the following positions (all based on the B-chain numbering shown in this Figure) are considered to be hydroxylated: A-chain, positions 11, 17, 23, 26, 35, 38, 47, 53, 59, 65, 80, 83 and 86: B-chain, positions 8, 11, 17, 26, 29, 32, 35, 38, 50, 56, 59, 65, 71, 74, 77, 80 and 83; C-chain, positions 8, 11, 14, 17, 26, 29, 38, 41, 47, 50, 56, 59, 68, 71, 74 and 80 (Reid, 1979). (*c*) Alignment of the *C*-terminal globular regions. – denotes that a gap has been inserted. The human B-chain numbering is used.

phage-like adherent cells are a primary source of mRNA for the chains and seem likely to be a major source of serum C1q, and this is consistent with studies carried out at the protein level (Tenner & Volkin, 1986; Bensa *et al.*, 1983). Although it might be expected that the chains would be expressed in equimolar amounts, Northern-blot analysis of polyadenylated mRNA isolated from a 7-day culture of adherent monocytes indicated that the concentration of the A-chain mRNA in monocyte-derived macrophages was several-fold higher than that of the B-chain or C-chain mRNAs.

Use of the A-, B- and C-chain probes in Southern-blot analysis of genomic DNA from C1q-deficient patients from three different families (B, EM and F, as detailed in the Materials and methods section) showed that none of these affected individuals possessed any gross deletions or insertions in their structural genes. Identical Southern-blotting patterns were obtained for DNA from the C1q-deficient patients, putative heterozygotes for the deficiency and normal individuals. To date, the only C1q deficiency that has been fully characterized at the DNA level is in a patient who showed the loss of a *Taq*I restriction site in the B-chain gene by virtue of a homozygous point mutation in the DNA coding for residue 150 in the B-chain that results in the appearance of a termination codon at this position (McAdam *et al.*, 1988). No material corresponding to the A- or C-chains or a truncated B-chain appeared to be present in the patient's serum, suggesting that a virtually complete B-chain must be synthesized before assembly and secretion of any modified form of C1q. *Taq*I was one of the restriction enzymes used in the study of the genomic DNA from the three C1q-deficient families examined in the present investigation, and no evidence for a similar point mutation to that characterized previously (McAdam *et al.*, 1988) was seen. Use of the human C-chain probe has revealed the lack of two *Sty*I restriction-enzyme sites in the C-chain gene in three C1q-deficient patients, from the same family, and at least one of these sites lies within the coding sequence for the C-chain (Petry *et al.*, 1990). Thus it seems likely that the lesions underlying the genetic deficiency of C1q are varied and may be attributable to a number of different point mutations, or small deletion or insertion events, either within the structural portion of the gene or in the control regions. The Southern-blotting data also provided no evidence for the existence of more than one type of gene coding for each chain, which is relevant to the finding that cultured fibroblasts, from C1q-deficient patients and a normal individual, could synthesize and secrete a C1q-like molecule having chains of apparently greater molecular mass than those of normal C1q (Skok *et al.*, 1981). This observation was interpreted as there being distinct genes for normal serum C1q. However, in view of the Southern-blotting data, it seems more likely that the production of C1q (or a C1q-like molecule) in fibroblasts would involve the same genes and that there would be fibroblast-specific and macrophage-specific controlling elements that would allow for differential expression of the chains of C1q between cell types.

DNA sequencing has allowed the complete derivation of the amino acid sequences of the A- and C-chains of human C1q (Figs. 1 and 2), which, along with the sequence for the B-chain (Reid, 1985), gives the entire amino acid sequence of the C1q molecule (Fig. 4). These three amino acid sequences are compared, in Fig. 4, with each other and with the derived sequence of the B-chain of mouse C1q (Wood *et al.*, 1988; Petry *et al.*, 1989). The derived sequence of the A-chain of mouse C1q has been established and presented (Petry & Loos, 1989), but the complete sequence has not yet been published. The human and mouse C1q chains all show hydrophobic leader peptides that are cleaved off and not found in the mature serum forms of the chains (Fig. 4*a*). The three human leader sequences, apart from the presence of a leucine-rich core in the B- and C-chains, do not show any significant degree of identity (Fig. 4*a*). The human A- and B-chain leader sequences show approx. 50% identity with the corresponding mouse A- and B-chain sequences. In each chain the amino acid occupying position −1 is either alanine or threonine, which conforms to the general rule that the amino acid in this position, which precedes cleavage by the signal peptidase, tends to be small and with an uncharged side chain (von Heijne, 1983). The complete amino acid sequences of the A- and B-chains, and the first 94 residues of the C-chain, of normal human serum C1q had previously been established by protein sequencing (Reid, 1979; Reid *et al.*, 1982). Comparison of the derived amino acid sequence of the human A-chain shown in Fig. 1 with the published protein data revealed five areas of discrepancy (as detailed in the legend to Fig. 1). Three of these differences were predicted from both cDNA and genomic sequences and all four changes increased the similarity shown between the human and mouse A-chains, and thus these discrepancies appear likely to be due to errors in amino acid sequencing rather than to polymorphism. The fifth discrepancy involves the *C*-terminal residues of the A-chain, which from the DNA sequencing are predicted to be -Leu-Ile-Phe-Pro-Ser-Ala. Again the derived amino acid sequence is taken to be correct in view of the fact that the protein data relied heavily on the use of

(*a*)

```
CX     T G M P V S A F T V I L S K A Y P - G A T V P I K F D K I L Y N R Q Q H Y D P R  539-578
CVIII  P A Y E M P A F T A E L T A P F P - P V G A P I K F D R L L Y N G R Q N Y N P Q  611-649
C1q B  K A T Q K I A F S A T R T I N V P L R R D Q T I R F D H V I T N M N N N Y E P R   90-129

CX     T G I F T C R I P G L Y Y F S Y H V H A K G T N V W V A L Y K N G S P V - - M Y  579-616
CVIII  T G I F T C E V P G V Y Y F A Y H V H C K G G N V W V A L F K N N E P V - - M Y  650-687
C1q B  S G K F T C K V P G L Y Y F T Y H A S S R G - N L C V N L M R G R E R A Q K V V  130-168

CX     T Y D E Y Q K G Y L D Q A S G S A V I D L M E N D Q V W L Q L P N S E S N G L Y  617-656
CVIII  T Y D E Y K K G F L D Q A S G S A V L L L R P G D R V F L Q M P S E Q A A G L Y  688-727
C1q B  T F C D Y A Y N T F Q V T T G G M V L K L E Q G E N V F L Q - - A T D K N S L L  169-206

CX     S S E Y V H S S F S G F L F A Q - I      657-673
CVIII  A G Q Y V H S S F S G Y L L Y P - M      728-744
C1q B  G M E G A N S I F S G F L L F P D M E A  207-226
```

(*b*)

	CX	CVIII	C1q A	C1q B
CVIII	29.8			
C1q A	12.2	12.0		
C1q B	12.7	13.2	17.5	
C1q C	12.6	12.9	17.2	24.9

Fig. 5. Comparison of type X (CX) and type VIII (CVIII) collagen NC1 domains, from chicken chrondrocytes and rabbit corneal endothelial cells (Yamaguchi *et al.*, 1989; Ninomiya *et al.*, 1986) respectively, with the globular head region of human C1q B-chain (C1q B) (Reid, 1985)

(*a*) Alignment of CX, CVIII and C1q B was made by the AMPS program (Barton & Sternberg, 1987). Gaps are inserted to maximize sequence similarities, identical residues present in all three proteins are boxed and conservative replacements are underlined. Residue numbering is indicated at the right of Figure. (*b*) Standard-deviation scores for pairwise alignment of CX, CVII, C1q B and equivalent regions of the C1q A and C-chains by using the AMPS program. A bias of 6 was added to each term of the mutation data matrix and a gap penalty of 6 was used with 100 random runs performed to establish mean random scores. Of the three chains in C1q, the B-chain has the highest degree of sequence identity with CX (34%) and CVIII (38%).

carboxypeptidases and it was difficult to monitor quantitatively the order of release of the proline/phenylalanine and isoleucine/leucine pairs of residues. Comparison of the derived amino acid sequence for the human C-chain also revealed differences with the published protein data (as detailed in the legend to Fig. 3), but it is not yet known if any of these differences represent polymorphic sites or if they are errors made in the protein sequencing.

Comparison of the mature forms of the human A-, B- and C-chain amino acid sequences shows that there are four conserved cysteine residues in each of the chains, at positions 4, 135, 154 and 171 (B-chain numbering; Fig. 4*c*), i.e. a total of $18 \times 4 = 72$ cysteine residues per molecule. The cysteine residues at position 4 in each chain (Fig. 4*b*) are involved in forming six A–B and three C–C disulphide-linked dimers, and these are the only inter-chain disulphide bonds in human C1q (Reid, 1976, 1983; Hughes-Jones & Gardner, 1979). Of the 54 cysteine residues present in the 'heads' formed from the globular regions of the A-, B- and C-chains 36 appear to be involved in intra-chain disulphide bonds and 18 appear to be present as free thiol groups (i.e. one free thiol group per chain, three free thiol groups per head). The value of 64 cysteine residues (26 disulphide bonds and 12 free thiol groups) as reported for C1q by Heusser *et al.* (1975) seems likely to be an underestimate in view of the value of 72 residues from the derived protein sequence (Fig. 4). The lack of a cysteine residue at position 171 in the mouse A-chain (Petry & Loos, 1989) suggests that the completely conserved cysteine residues at positions 135 and 154 seen in the sequences available (Fig. 4*c*) may be involved in formation of an intra-chain disulphide bond in each chain whereas Cys-171 may contribute a free thiol group when present. It has been proposed that the free thiol groups in C1q can be involved in the formation of disulphide bonds between C1q and IgG when the IgG is bound to Sepharose or present in immune complexes (Martin *et al.*, 1990), thus providing a potentially important physiological role for these free thiol groups.

Alignment of the *C*-terminal regions of the chains of human C1q and mouse C1q indicates that there is a 'framework' of conserved, predominately hydrophobic and neutral, amino acid residues (approx. 25% of the total residues in each globular region). These residues are probably responsible for maintaining the structure of the 'heads' of the C1q molecule rather than being involved in specific ionic binding of C1q to the Fc regions of IgG or IgM. The amino acid sequences of the non-collagen-like portions of the C1q chains (from residue 90 to the *C*-terminus; Fig. 4*c*) show a strong similarity to the *C*-terminal regions of type VIII (Yamaguchi *et al.*, 1989) and type X (Ninomiya *et al.*, 1986) collagens (Fig. 5). The alignment scores of 12.0 and 13.2 for comparison of residues 90–226 (B-chain numbering) of the chains of C1q with the *C*-terminal regions of type VIII and type X collagen illustrate the highly significant degree of sequence similarity between the three proteins, since scores of above 6.0 are considered very significant. The presence of many completely conserved hydrophobic residues is suggestive

that these are framework residues required to build a compact globular structure, although in the case of C1q this is formed from three different chains whereas the globular 'heads' of type VIII and type X collagens are considered to be formed as homotrimers. One possible role of this highly conserved structure may be to allow the correct alignment of the Gly-Xaa-Yaa repeating triplets so that efficient triple-helix formation takes place. The residues in C1q that lie outside the regions of similarity between C1q and type VIII and X collagens may be involved in C1q function, i.e. the binding of the globular 'heads' to the Fc regions of IgG and IgM, by ionic interaction with a motif such as -Glu_{318}-Xaa-Lys_{320}-Xaa-Lys_{322} found in the Cγ2 domain of IgG (Duncan & Winter, 1988). No specific binding functions have been attributed to the relatively basic *C*-terminal globular domains of the type VIII and type X collagens, which are considered to interact with polyanions and perhaps play structural roles in the formation of Descement's membrane and the development of cartilage respectively. However, it is clear that this sequence of approx. 130 residues (Fig. 5), present in all three proteins, defines a new type of structural domain.

In chemical and electron-microscopy studies, three other molecules have been found to be structurally similar to C1q with respect to being composed of chains linked at the *N*-terminal end within a short stretch of non-collagen-like amino acid sequence, followed by distinct collagen-like and then globular-type regions. These three proteins, mannan-binding protein (Drickamer *et al.*, 1986; Lu *et al.*, 1990), lung surfactant protein SP-A (Voss *et al.*, 1988) and conglutinin (Davis & Lachman, 1984; Strang *et al.*, 1986), are lectins, and their globular domains belong to the Ca^{2+}-dependent C-type lectin consensus amino acid sequence (as defined by Drickamer, 1988) and therefore show no similarity in sequence to the globular *C*-terminal regions of C1q. However, lung surfactant protein SP-A is virtually indistinguishable from C1q in the electron microscope, both displaying the very characteristic hexameric structure described for C1q in the Introduction. Mannan-binding protein is similar to C1q and lung surfactant protein SP-A in overall structure but is found mainly in the form of trimers and tetramers with relatively little of the hexameric form. It is the hexameric form of mannan-binding protein that has recently been demonstrated as capable of activating the $C1r_2C1s_2$ proenzyme complex of the classical pathway of complement without the involvement of C1q (Lu *et al.*, 1990), and this may be an important antibody-independent route of complement activation in certain infections where there has been no or little antibody response. Conglutinin has a more flexible 'spider-like' structure when viewed in the electron microscope. C1q, lung surfactant protein SP-A and mannan-binding protein all have a characteristic 'bend' approx. half-way along their collagen-like triple-helical regions, which corresponds to one or more interruptions in the repeating nature of the Gly-Xaa-Yaa triplets, whereas conglutinin does not appear to show this feature. In human C1q A-chain a threonine residue is inserted between two triplets at position A-39, and in the C-chain an alanine residue is substituted for a glycine residue at position C-36, whereas the B-chain does not have an interruption but does appear to possess an extra triplet (Figs. 4*b* and 6). Modelling studies have placed all the Gly-Xaa-Yaa triplets in C1q in an almost perfect triple helix, with the helix 'bending' in the position of the interruptions (Kilchherr *et al.*, 1985). The human C1q A-, B- and C-chain genes each contain a single intron (Fig. 2), and in each case the intron/exon boundary lies between the first and second bases for the codon for a glycine residue just before the interruption of the Gly-Xaa-Yaa repeating sequences for the A- and C-chains, and over the equivalent region in the B-chain. In lung surfactant protein SP-A a cysteine residue is found in a 'glycine' position in the 14th triplet and a proline residue is inserted between the 13th and 14th triplets (Fig. 6). In mannan-binding protein an interruption is seen in the form of a deletion of one residue in the eighth triplet. Genomic sequencing (White *et al.*, 1985; Sastry *et al.*, 1989; Taylor *et al.*, 1989) has shown that in both lung surfactant protein SP-A and mannan-binding protein there is an intron located within a glycine codon located just before the positions where the interruptions to the collagen-like sequence are seen at the protein level (Fig. 6). It therefore seems possible that the positioning of the intron within the genes of these proteins may confer some retention of structurally important information, i.e. to ensure that the triple-helical regions exhibit a 'characteristic' bend at a precise position within the protein. In addition to C1q, mannan-binding protein and lung surfactant protein, the human type IV collagen also shows this feature at one of the numerous short interruptions seen in its collagen-like region (Soininen *et al.*, 1987; Fig. 6).

A chain:	G E P **G** A P G I R T G .
B chain:	G E K **G** L P G L A - G .
C chain:	G E P **G** I P A I K - G .
MBP:	G E P **G** Q - G L R . G .
SP-A:	G P P **G** P M G P P G E M P C P P G .
Type IV collagen:	G E P **G** L P G A - - G

Fig. 6. Sequence similarity between the chains of proteins containing collagen-like sequences at the point where there is interruption to the collagen-like sequence

Interruptions are seen in the collagen-like sequences of the human C1q A- and C-chains, the human mannan-binding protein (MBP) (Taylor *et al.*, 1989), the human lung surfactant protein SP-A (White *et al.*, 1985; Sastry *et al.*, 1989) and human type IV collagen (Soininen *et al.*, 1987). In each case an intron is found within the codon for the glycine (in **boldface**) located between two and nine residues before the interruption to the collagen-like sequence. The human B-chain does not contain an interruption in its collagen-like sequence but does show the presence of an intron within a codon for glycine at an equivalent position to that found for the A- and C-chains.

The collagen-like regions of C1q and mannan-binding protein interact with the $C1r_2C1s_2$ complex and also the cell-surface C1q receptor via ionic bonds (Reid *et al.*, 1977; Siegel & Schumaker, 1983; Malhotra *et al.*, 1990). Thus, in view of the similarity in binding properties mediated via their collagen-like regions, it might be expected that the chains of C1q and mannan-binding protein might show some conservation of charged residues over their collagen-like regions. Alignment of the amino acid sequences corresponding to the collagen-like regions of the chains of C1q shows only limited similarity between the chains apart from the presence of the repeating glycine residues in the Gly-Xaa-Yaa triplets (Fig. 4*b*). Only the cysteine residue at position 4, three proline residues at positions 17, 59 and 67 and a glutamic acid residue at position 34 are completely conserved among the three human C1q chains and also in human mannan-binding protein (Taylor *et al.*, 1989). This conserved glutamic acid residue is within the Gly-Xaa-Yaa triplet immediately preceding the region where the chains of both human and mouse C1q and human mannan-binding protein show an interruption in the Gly-Xaa-Yaa repeating nature of the collagen-like sequence. It is therefore possible that both this negatively charged glutamic acid residue plus the 'bends' seen in triple-helical regions are important structural features of the interaction of C1q and mannan-binding protein with the $C1r_2C1s_2$ complex and/or the C1q receptor found on a wide range of lymphoid cells.

We thank Dr. A. J. Day for carrying out the sequence alignment studies using the AMPS program.

REFERENCES

Barton, G. J. & Sternberg, M. J. E. (1987) J. Mol. Biol. **198**, 327–337

Bensa, J. C., Reboul, A. & Colomb, M. G. (1983) Biochem. J. **216**, 385–392

Brodsky-Doyle, B., Leonard, K. R. & Reid, K. B. M. (1976) Biochem. J. **159**, 279–286

Chapuis, R. M., Hauptmann, G., Grosshaus, E. & Isliker, H. (1982) J. Immunol. **129**, 1509–1512

Chirgwin, J. M., Przbyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) Biochemistry **18**, 5294–5299

Davis, A. E. & Lachman, P. J. (1984) Biochemistry **23**, 2139–2144

Drickamer, K. (1988) J. Biol. Chem. **263**, 9557–9560

Drickamer, K., Dordal, M. S. & Reynolds, L. (1986) J. Biol. Chem. **261**, 6878–6887

Duncan, A. R. & Winter, G. (1988) Nature (London) **332**, 738–740

Frischauf, A. M., Lehrach, H., Poustka, A. & Murray, N. J. (1983) J. Mol. Biol. **170**, 827–842

Glover, D. M. (1985) DNA Cloning, vol. 1, pp. 49–78, IRL Press, Oxford

Grosshaus, E., Hauptmann, G. & Friedel, J. (1987) Pediatr. Dermatol. News **6**, 115–120

Hanahan, D. & Meselson, M. (1983) Methods Enzymol. **100**, 333–342

Heusser, C. H., Boesman, M., Knobel, H. R., Jacot-Guillarmod, H. & Isliker, H. (1975) Immunochemistry **12**, 213–219

Hughes-Jones, N. C. & Gardner, B. (1979) Mol. Immunol. **16**, 697–701

Kaufman, K., Snider, J. V., Spurr, N. K., Schwartz, C. E. & Sodetz, J. M. (1989) Genomics **5**, 475–480

Kilchherr, E., Hofmann, H., Steigemann, W. & Engel, J. (1985) J. Mol. Biol. **186**, 403–415

Knobel, H. R., Villiger, W. & Iskiker, H. (1975) Eur. J. Immunol. **5**, 78–81

Knott, V., Rees, D. J. G., Cheng, Z. & Brownlee, G. G. (1988) Nucleic Acids Res. **16**, 2601–2612

Lu, J., Thiel, S., Wiedemann, H., Timpl, R. & Reid, K. B. M. (1990) J. Immunol. **144**, 2287–2295

Malhotra, R., Thiel, S., Reid, K. B. M. & Sim, R. B. (1990) J. Exp. Med. **172**, 955–959

Maniatis, T., Fritsch, E. P. & Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor

Martin, H., Kaul, M. & Loos, M. (1990) Eur. J. Immunol. **20**, 1641–1646

Maxam, A. M. & Gilbert, W. (1980) Methods Enzymol. **65**, 499–560

McAdam, R. A., Goundis, D. & Reid, K. B. M. (1988) Immunogenetics **27**, 259–264

Ninomiya, Y., Gordon, M., van der Rest, M., Schmid, T., Lisenmayer, T. & Olsen, B. R. (1986) J. Biol. Chem. **261**, 5041–5050

Pardo-Manuel, F., Rey-Campos, J., Hillarp, A., Dahlback, B. & Rodriguez de Cordoba, S. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 4529–4532

Petry, F. & Loos, M. (1989) Complement Inflammation **6**, 387

Petry, F., Reid, K. B. M. & Loos, M. (1989) FEBS Lett. **258**, 89–93.

Petry, F., Starsia, Z., Reid, K. B. M., Sellar, G. C. & Loos, M. (1990) Complement Inflammation **7**, 117

Reid, K. B. M. (1976) Biochem. J. **155**, 5–17

Reid, K. B. M. (1979) Biochem. J. **179**, 367–371

Reid, K. B. M. (1983) Biochem. Soc. Trans. **11**, 1–12

Reid, K. B. M. (1985) Biochem. J. **231**, 729–735

Reid, K. B. M. & Porter, R. R. (1976) Biochem. J. **155**, 19–23

Reid, K. B. M. & Thompson, E. O. P. (1978) Biochem. J. **173**, 863–868

Reid, K. B. M., Sim, R. B. & Faiers, A. P. (1977) Biochem. J. **161**, 239–245

Reid, K. B. M., Gagnon, J. & Frampton, J. (1982) Biochem. J. **203**, 559–569

Rey-Campos, J., Rubinstein, P. & Rodriguez de Cordoba, S. (1988) J.Exp. Med. **167**, 664–669

Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467

Sastry, K., Herman, G. A. & Day, L. (1989) J. Exp. Med. **170**, 1175–1189

Sellar, G. C. & Reid, K. B. M. (1989) Complement Inflammation **6**, 399

Sellar, G. C., Goundis, D., McAdam, R. A., Solomon, E. & Reid, K. B. M. (1987) Complement **4**, 225

Shelton, E., Yonemasu, K. & Stroud, R. M. (1972) Proc. Natl. Acad. Sci. U.S.A. **69**, 65–68

Siegel, R. C. & Schumaker, V. N. (1983) Mol. Immunol. **20**, 53–66

Skok, J., Solomon, E., Reid, K. B. M. & Thompson, R. A. (1981) Nature (London) **292**, 549–551

Soininen, R., Haka-Risku, T., Prokop, D. & Tryggvason, K. (1987) FEBS Lett. **225**, 188–194

Southern, E. M. (1979) Methods Enzymol. **68**, 152–176

Strang, C. J., Slayter, H. S., Lachmann, P. J. & Davis, A. E. (1986) Biochem. J. **234**, 381–389

Svehag, S. E., Manheim, L. & Bloth, B. (1972) Nature (London) New Biol. **238**, 117–118

Taylor, M. E., Bricknell, P. M., Craig, R. K. & Summerfield, J. A. (1989) Biochem. J. **262**, 763–771

Tenner, A. J. & Volkin, D. B. (1986) Biochem. J. **233**, 451–458

von Heijne, G. (1983) Eur. J. Biochem. **133**, 17–21

Voss, T., Eistetter, H., Schafer, K. P. & Engel, J. (1988) J. Mol. Biol. **201**, 219–227

Wahl, G. M., Stern, M. & Stark, G. (1979) Proc. Natl. Acad. Sci. U.S.A. **76**, 3683–3687

White, R. T., Damm, D. & Miller, J. (1985) Nature (London) **317**, 361–363

Wood, L., Pulaski, S. & Vogeli, G. (1988) Immunol. Lett. **17**, 115–120

Yamaguchi, N., Benya, P. D., van der Rest, M. & Nimomiya, Y. (1989) J. Biol. Chem. **264**, 16022–16029

Received 21 June 1990/17 August 1990; accepted 21 August 1990